

Running head: Grouping Survey Items and Discriminant Validity

Discriminant Validity Where There Should Be None: Positioning Same-Scale Items in Separated Blocks of a Questionnaire

Accepted for publication in Applied Psychological Measurement

DOI: 10.1177/0146621614531850

BERT WEIJTERS^{1,*}, ALAIN DE BEUCKELAER^{2,1,3}, and HANS BAUMGARTNER⁴

¹*Ghent University, Department of Personnel Management, Work and Organizational Psychology, Belgium.*

²*Renmin University of China, School of Sociology and Population Studies, China.*

³*Radboud University Nijmegen, Institute for Management Research, the Netherlands.*

⁴*Smeal College of Business, The Pennsylvania State University, U.S.*

Contact information:

*Author for correspondence: Bert Weijters, Ghent University, Department of Personnel Management, Work and Organizational Psychology, Dunantlaan 2, B-9000 Ghent, Belgium. Tel: +32 9 264 62 96. E-mail: bert.weijters@ugent.be. The data-file is available upon request from the first author.

Alain De Beuckelaer: Radboud University Nijmegen, Institute for Management Research, Thomas van Aquinostraat 3, 6525 GD Nijmegen, The Netherlands, E-mail: A.DeBeuckelaer@fm.ru.nl

Hans Baumgartner: Smeal College of Business at The Pennsylvania State University, Department of Marketing, 482 Business Building, University Park, PA 16802, United States. Tel.: +1 814 863 3559; fax: +1 814 865 3015. E-mail: hansbaumgartner@psu.edu

Abstract.

In questionnaires items can be presented in a grouped format (same-scale items are presented in the same block) or in a randomized format (items from one scale are mixed with items from other scales). Some researchers have advocated the grouped format because it enhances discriminant validity. The current study demonstrates that positioning items in separate blocks of a questionnaire may indeed lead to increased discriminant validity, but this can happen even in instances where discriminant validity should not be present. In particular, we show that splitting an established unidimensional scale into two arbitrary blocks of items separated by unrelated buffer items results in the emergence of two clearly identifiable but artificial factors that show discriminant validity.

Key Words:

Confirmatory factor analysis; discriminant validity; grouping (survey) items; inter-item correlation; randomizing (survey) items; response styles; survey methods.

Author note:

The data-file is available upon request from the first author.

Survey data are a key source of information for social scientists and, consequently, the validity of questionnaire-based measurement is of major concern to researchers in psychology. Previous research has shown that the way in which items and scales are positioned in a questionnaire can affect the psychometric properties and validity of the measures used (Schriesheim, Solomon, & Kopelman, 1989b). Studying the effects of item positioning in questionnaires is particularly interesting because item positioning provides an efficient means of improving measurement quality (Schriesheim & DeNisi, 1980)

One important question related to the positioning of items and scales is whether to group items belonging to the same multi-item scale into one block (i.e., grouped format) or whether to randomly distribute them over the questionnaire, interspersed with items belonging to other scales (i.e., randomized format). Traditionally, and particularly so in the personality area, researchers have tended to recommend and use randomized item arrangements (Franke, 1997). The primary motivation for this practice seems to be the desire to minimize demand effects and faking because randomization makes it less transparent what is being measured (McFarland, Ryan, & Ellis, 2002). However, randomization comes at a cost, because randomized item arrangements require greater ability and motivation on the part of respondents (Schriesheim & DeNisi, 1980). Furthermore, since the relatedness of same-scale items is less obvious to respondents, randomization may also affect discriminant validity (Harrison & McLaughlin, 1996).

When it comes to measurement, discriminant validity is generally conceived as an important goal for researchers (Lehmann, 1988). With regard to item positioning, there is reason to believe that a grouped format enhances discriminant validity, although research that explicitly addresses this question is relatively scarce. Initial studies by Schriesheim and colleagues yielded somewhat inconclusive results (Schriesheim, 1981b; Schriesheim & DeNisi, 1980; Schriesheim, Solomon, & Kopelman, 1989a; Schriesheim et al., 1989b).

However, a subsequent influential study by Harrison and McLaughlin (1996) led to the recommendation that same-scale items should be grouped together in order to enhance discriminant validity. The current study demonstrates that evidence of discriminant validity realized by grouping same-scale items should be regarded with caution. The reason is that item grouping will tend to generate some level of discriminant validity even where there should be none. That is, indicators of discriminant validity are prone to artifacts caused by item positioning. In particular, we will demonstrate that randomly grouping items from an established unidimensional scale into two blocks separated by unrelated buffer items results in two artificial factors that show discriminant validity.

1. Literature review: grouping items and discriminant validity

In this section, we discuss existing research that addresses the question of how grouping versus randomizing items in a questionnaire affects discriminant validity. Schriesheim and DeNisi (1980) studied two leadership instruments with four dimensions each. The items were presented either in eight grouped and labeled sections or in random order. Based on a traditional multitrait-multimethod analysis, the authors concluded that the randomized format resulted in better discriminant validity. In a re-analysis of the same data, Schriesheim (1981a) used a hierarchical factor-analytic procedure to show that there was less communality (and thus greater discrimination) between the dimensions for the randomized format. In another analysis, Schriesheim (1981b) found that discriminant validity in the grouped format improved after controlling for leniency bias. As was pointed out by the authors themselves in subsequent research (Schriesheim et al., 1989a, 1989b), these conclusions may have limited generalizability as the same small sample ($N = 60$) was used for repeated analyses and a specific leadership instrument prone to leniency bias was employed.

Schriesheim et al. (1989a) and Schriesheim et al. (1989b) analyzed two data sets (in both papers the same two data sets were analyzed, but more advanced confirmatory factor analysis procedures were used in the second paper) in an effort to investigate the effect of item positioning on discriminant validity (among other things). In study 1, Schriesheim et al. (1989a) investigated four satisfaction and five job characteristics scales. Respondents ($N = 80$) completed both the grouped format (in which the scales were also labeled) and the random format of all scales in the same session (so the grouped/random format was a within-subject manipulation); presentation order was randomized between-subjects. Based on the traditional multitrait-multimethod approach (with grouped versus randomized format representing the distinct categories of the method factor), the authors argued that the grouped condition showed slightly better discriminant validity. In study 2 ($N = 80$), Schriesheim et al. (1989a) administered the same instruments used in study 1 repeatedly (5 to 8 weeks apart) to the same respondents (with time used as a method factor). Half of the participants received a randomized satisfaction instrument and a grouped job characteristics instrument in both sessions, the other half received a grouped satisfaction instrument and a randomized job characteristics instrument. The authors argued that the grouped format had a slight advantage in terms of discriminant validity as it showed somewhat better discrimination over time (i.e., distinct subdimensions of the scale correlated less strongly across the two scale administrations in the grouped format).

As part of a broader research program on cognitive effects of questionnaire design, Harrison and McLaughlin (1996) also investigated the effect of grouped versus randomized formats on discriminant validity. In the study of interest (study 2, $N = 392$), they used six different scales and two different response formats (Likert scales and easy/difficult ratings). Items were presented in boxes. In the grouped format, items in the same box were from the same scale. In the randomized format, items in the same box were from different scales. Each

respondent saw both grouped and random sets. The authors found better discriminant validity in the grouped condition and concluded that “physically grouping items on a questionnaire slightly enhances internal consistency and discriminant validity, by enhancing the within-set commonalities and between-set distinctions that guide respondents to retrieve relevant caches of information” (p. 329). They therefore suggested that the grouped format should be preferred over the randomized format. This recommendation has been taken to heart by researchers who have constructed and validated scales in recent years and who have used a grouped format, partly in order to enhance discriminant validity (Conte, Dean, Ringenbach, Moran, & Landy, 2005; Shipp, Edwards, & Lambert, 2009).

Before it becomes accepted practice to group items in an effort to improve discriminant validity, an important note of caution is in order. Although previous research has demonstrated that a grouped presentation order (where all the items belonging to the same scale form one item block) enhances discriminant validity, the beneficial effects of grouping may be a methodological artifact. One way to demonstrate this is to show that grouping items that belong to the same multi-item scale into separate blocks may result in discriminant validity *even when there should be none*. In particular, if items from an established unidimensional scale are arbitrarily put into separate blocks of items, a multidimensional factor structure consisting of several distinct factors that exhibit discriminant validity may emerge, even though there is no substantive validity to the separate factors.

In the current study we therefore investigate what happens when items from one and the same scale are included in a questionnaire as if they were measuring two separate constructs. We position half of the items in the scale at the beginning of the questionnaire, the other half at the end, with unrelated filler items in between. Our hypothesis is that this positioning will yield two artificial factors (whereas all items should normally load on one underlying factor if positioning had no influence on the factorial structure). In practice,

researchers will typically not split up a scale into two or more blocks, unless the scale contains many items and the items have to be presented on multiple pages or multiple screens in online surveys (e.g., in an effort to avoid the need for scrolling). And even if it is necessary to divide the entire scale into multiple blocks, the blocks would not be separated by unrelated filler items. However, we want to investigate whether an admittedly strong blocking manipulation can yield two separate factors when there should only be one factor based on the common content of the items. Our approach diverges from earlier work on grouping of items, in which items were grouped into blocks based on common content within blocks and supposedly different content in different blocks. Unfortunately, such a design confounds desirable discriminant validity (due to lack of overlap in content) with undesirable discriminant validity (due to positioning items from different scales in different blocks). The current study avoids this confound and shows that blocking of items can result in artificial and undesirable discriminant validity.

2. Method

We collected data from the online U.S. panel of a global panel provider. In our sample ($N = 523$), 51.1% is female, age ranges from 20 to 75 years ($M = 45.1$, $SD = 14.4$), and 79.3% of respondents enjoyed at least some college education. The questionnaire contained the scale of interest for the experimental format manipulation (discussed in more detail below) and 32 filler items. The 32 filler items consisted of eight four-item scales with diverse content, each shown on one page (i.e., in a grouped format). The filler items (i.e., the eight four-item scales) were included for a different study, unrelated to our current purposes, and all four-item filler scales were full scales if the original scale consisted of four items, or shortened versions of the original published multi-item scales otherwise. Representative items from some of the scales used include “I enjoy tasks that require me to be exact” (Need for Precision), “When I see a new or different brand on the shelf, I often pick it up just to see what it is like” (Shopping

Innovativeness), and “I feel satisfied with the way my body looks right now” (Self-esteem, Appearance Dimension). All items were rated on a five-point Likert format with response categories labeled strongly agree, agree, neither agree nor disagree, disagree, strongly disagree (the response options were not visibly numbered but for analysis they were coded with consecutive integers from 1=strongly agree to 5 = strongly disagree).

As the scale of interest for the experimental format manipulation, we selected an eight-item frugality scale (Lastovicka, Bettencourt, Hughner, & Kuntze, 1999). Lastovicka et al. (1999) reported Cronbach’s alpha coefficients of .85 and .87 in two random halves of a nonstudent adult quota sample ($N = 213$), with χ^2 (20) values of 25.93 ($p = .17$) and 31.92 ($p = .04$), respectively, for a one-factor model in which factor loadings ranged from .53 to .77. In two subsequent convenience samples ($N = 57$, and $N = 90$), coefficient alphas of .88 and .73 were obtained. In a final sample ($N = 164$, from a mail survey of randomly selected adults), coefficient alpha was .80 and a one-factor model had an overall fit of χ^2 (20) = 30.93 ($p = .06$), with factor loadings ranging from .63 to .73. Rick, Cryder, and Loewenstein (2008) reported a coefficient alpha of .84 in a large convenience sample ($N = 1,955$). In sum, in past research the frugality scale fit statistics supported a one-factor model and internal consistency was found to be high.

The frugality scale (Lastovicka et al., 1999) consists of eight items all coded in the same direction, so no artificial subfactors can emerge due to item reversals (Emons, 2009). Equal directionality of all items makes it convenient to split the scale into two random blocks of four items. Four frugality items were shown on one screen before the filler items (in block 1), and four were shown after the filler items (in block 2). We randomly assigned respondents to two experimental conditions. In the first condition ($N = 268$), frugality items 1 to 4 were assigned to block 1, and items 5 to 8 were assigned to block 2. In the second condition ($N = 255$), items 3, 4, 5, 6 were included in block 1, whereas items 1, 2, 7 and 8 were included in

block 2. Figure 1 shows the factor model that is hypothesized to fit the data best as a result of the grouping manipulation for each condition. Within the two blocks, item order was randomized across respondents (e.g., when items 1, 2, 3, and 4 were in block 1, the four items were shown in random order to different respondents); the reason is that we are mainly interested in the effect of the grouping manipulation, not order effects within blocks. Table 1 reports the items contained in the scale and univariate descriptive statistics per condition.

<Insert Figure 1 about here>

<Insert Table 1 about here>

3. Results

To assess the impact of the grouping manipulation, we report two sets of analyses. First, we study the influence of item grouping on the inter-item correlations directly, which makes it possible to quantify the effect on item correlations of positioning same-scale items in separated blocks rather than a single block. Second, we assess the factor structures that emerge from the correlations in the various conditions. This allows us to evaluate the relative fit of one- and two-factor models and to assess whether a two-factor model resulting from same-scale items being positioned in different blocks indeed attains artificial discriminant validity based on commonly used criteria for ascertaining discriminant validity.

3.1. Effect on correlations

Table 2 displays the raw inter-item correlations for the two conditions, for a total of 56 correlations (28 unique correlations per condition), as well as the expected correlations based on Lastovicka et al. (1999). We obtained the expected correlations as follows. First, we averaged the factor loadings for each of the eight items across the three studies reported in Table 1 of Lastovicka et al. (1999). To the best of our knowledge, only these three studies

provide factor loading estimates for this scale. Then, for each item pair, we computed the expected correlation as the product of the factor loadings of the two items involved (as the factor loadings presumably come from completely standardized solutions). So, for example, item 1 has a mean factor loading of .72, item 2 has a mean factor loading of .65, resulting in an expected correlation of $.72 * .65 = .47$. The expected correlation matrix serves as a baseline, since it is derived from scale administrations in which the items are positioned in the same block, without unrelated items interspersed between them.

<Insert Table 2 about here.>

To test how the grouping manipulation affects the inter-item correlations, we estimate a two-level regression model using the observed correlations in conditions 1 and 2 (as reported in Table 2) as the dependent variable. Single-level regression models with correlations as the dependent variable have been used previously in similar research (Lehmann, 1988; Weijters, Geuens, & Schillewaert, 2009), but due to our nested data structure (i.e., two repeated observations for each item pair), we rely on a two-level regression model. For each item pair, we treat the two observed correlations corresponding to conditions 1 and 2 as replications (this is the level-1 model), and there are 28 distinct item pairs (this is the level-2 model). To clarify the design, we have 28 independent correlations (one for each unique item pair). The 28 correlations are observed twice (i.e., in two different samples), some under identical conditions (i.e., in both samples the two items in the pair were either in the same block or in separated blocks) and others under different conditions (i.e., once in the same block and once in a separated block). Given this data structure, we specify the following model:

$$r_{ic} = \beta_{0i} + \beta_1 * SEPARATED_{ic} + \beta_2 * CONDITION2_{ic} + \varepsilon_{ic} \quad (1)$$

$$\beta_{0i} = \gamma_0 + \gamma_1 * \rho_i + \zeta_i \quad (2)$$

In the level-1 model (equation 1), r_{ic} is the correlation for item pair i (where i refers to one of the 28 item pairs) in condition c (i.e., condition 1 or condition 2), $SEPARATED_{ic}$ is a dummy variable taking on a value of zero if the items in the pair are positioned in the same block and a value of one if the items in the pair are positioned in different blocks, and $CONDITION2_{ic}$ is a dummy variable taking on a value of zero if r_{ic} comes from condition 1 and a value of one if r_{ic} comes from condition 2. The coefficient β_1 captures the difference in correlation between two items when they are positioned in two separated blocks rather than in the same block. Coefficient β_1 is of primary interest in our analysis, and the sign of the coefficient is predicted to be negative. The variable $CONDITION2_{ic}$ is included as a control variable in equation 1, and the coefficient β_2 captures potential differences in correlations between conditions 1 and 2. The intercept of equation 1, β_{0i} , is specified as a random coefficient that is allowed to vary across item pairs. Finally, ε_{ic} is the residual term in the level-1 model.

Equation 2 is the level-2 model according to which the level-1 random intercept, β_{0i} , is a function of ρ_i , the expected correlation for item pair i based on the factor loadings reported in Lastovicka et al. (1999) (see our earlier explanation). If, after controlling for $SEPARATED_{ic}$ and $CONDITION2_{ic}$, the inter-item correlations observed in our study are close to the correlations implied by the factor loadings reported in Lastovicka et al. (1999), the level-2 intercept γ_0 should be close to zero and the effect of ρ_i on β_{0i} , γ_1 , should be close to one (i.e., there should be an approximate one-to-one correspondence between the two sets of correlations). The residual term in the level-2 models is denoted ζ_i .

<Insert Table 3 about here.>

Table 3 reports the parameter estimates for the two-level model. In the level-2 model, the regression coefficient γ_1 , linking the observed correlations with the expected correlations, is significant and close to one, with a 90% confidence interval including one.

Correspondingly, a Wald χ^2 test, assessing whether β_2 differs significantly from one, is not significant ($\chi^2(1) = .220, p = .639$). The residual variance, $\text{var}(\zeta_i)$, is small though significant, and the intercept term γ_0 is close to zero and not significant. These parameter estimates show that, after controlling for the effects of $\text{SEPARATED}_{\text{IC}}$ and $\text{CONDITION2}_{\text{IC}}$, the observed correlations found in our study are consistent with the correlations that would be expected based on the factor loadings reported by Lastovicka et al. (1999). This finding may also be interpreted as an additional cross-validation of the frugality scale.

In the level-1 model, the effect of the $\text{SEPARATED}_{\text{IC}}$ dummy variable (β_1) is significant and negative, indicating a substantial drop in inter-item correlations ($\Delta r = -.165$) when two items are positioned in separated blocks rather than in the same block. Somewhat unexpectedly, the correlations are lower in condition 2 than in condition 1, as shown by the significant estimate for β_2 . A detailed look at the correlations reported in table 2 indicates that particularly the same-block correlations involving item 2 are surprisingly low in condition 2 (i.e., the correlations i1-i2, i2-i7 and i2-i8). Closer scrutiny of the item content shows that in condition 2, item 2 happened to be randomly placed with three other items that all refer to saving money (whereas item 2 does not refer to saving money), and this may have caused an accidental contrast effect that does not occur when all the items in the complete scale are presented together. Thus, putting items in random sub-blocks may sometimes inadvertently generate idiosyncratic context effects, in addition to the emergence of two artificial factors that we hypothesized and observed.

3.2. Effect on factor structures

To investigate how the grouping manipulation impacts the factor structure of a scale, we ran confirmatory factor analyses in Mplus version 6.11. As a baseline, we tested a simple model using the pooled (i.e., all) data from conditions 1 and 2, which assumes that all eight

items load on one common factor. This model fit the data badly, as manifested by a large χ^2 statistic, low CFI and TLI values, and high SRMR and RMSEA values (Hu & Bentler, 1999): $\chi^2(20) = 297.94, p < .001$; CFI = .815, TLI = .741, SRMR = .072, RMSEA = .163. To test whether conditions differing in how items are grouped lead to the emergence of two artificial factors, we also fitted the hypothesized two-factor models to the data for each condition separately (see Figure 1). Table 4 shows that, both in condition 1 and in condition 2, the hypothesized model fits the data well (see Table 4, first line per condition; the factors' composite reliabilities in this model range from .71 to .86).

<Insert Table 4 about here.>

To more conclusively demonstrate that the two-factor model corresponding to the two manipulated item blocks fits the data better than alternative two-factor structures, we evaluated all possible permutations of alternative two-factor models with four items per factor (there are 35 unique permutations). To illustrate, one such permutation would have items 1, 4, 5 and 7 load on factor 1, and items 2, 3, 6 and 8 load on factor 2. Figure 2 displays the histogram of the χ^2 values for all these models. The χ^2 values for the hypothesized models in conditions 1 and 2 (see Figure 1) are the two observations on the left-hand side in the figure, shown in the circle. Figure 2 clearly demonstrates that, in both conditions and in terms of χ^2 , the hypothesized two-factor model outperforms all alternative two-factor models (all models have the same degrees of freedom, so the χ^2 values are directly comparable).

<Insert Figure 2 about here.>

Table 4 reports the fit indices averaged across all permutations of alternative two-factor models, as well as the best fitting model among these alternative two-factor models. Once more, it can be concluded that the hypothesized two-factor model outperforms the best

fitting alternative two-factor model. Table 4 also includes the fit indices for a one-factor model, fitted to the data in conditions 1 and 2 separately. Once more this model fits poorly.

Based on the model estimates of the two-factor model (as estimated in each condition separately), we also assessed the discriminant validity of the two emergent factors. Two tests of discriminant validity are commonly reported. First, if two factors are distinct, their factor correlation should be different from one. Second, Fornell and Larcker (1981) proposed a stronger test of discriminant validity. The intuition is that a construct should have more overlap with its own indicators than with a supposedly distinct other construct. Shared overlap between constructs, or shared variance (SV), is simply the squared factor correlation, and shared overlap between a construct and its indicators can be indexed by average variance extracted (AVE), which is defined as the average of the shared variances between a construct and its indicators. Operationally, AVE equals the sum of the squared factor loadings multiplied by the factor variance divided by the total variance of all the items (i.e., the sum of the squared factor loadings multiplied by the factor variance plus the sum of the residual variances). Alternatively, AVE can also be computed as the average squared correlation between each of the items and the underlying factor. The criterion proposed by Fornell and Larcker (1981) is that SV should be smaller than the AVE for each of the factors for which discriminant validity is to be assessed. The inter-factor correlations (and their related 90% confidence intervals) are .61 (.53, .70) in Condition 1 and .61 (.52, .70) in Condition 2. The factor correlations are clearly smaller than 1, as the confidence intervals around the estimated factor correlations have upper boundaries substantially smaller than one. Furthermore, in both Condition 1 and Condition 2, the shared variance between the two factors (SV) is smaller than the AVE for the two factors. In particular, the SV is .38 in both conditions, whereas the AVE for factor 1 and factor 2 is .54 and .61 in Condition 1, and .53 and .40 in Condition 2.

Thus, the analysis provides evidence strongly supporting discriminant validity for two artificial factors.

4. Discussion

In the current study, we investigated the impact of grouping items in questionnaires. In previous work on grouping items, the items that were grouped together formed one complete scale. Consequently, within-block convergence and between-block divergence were favorable outcomes in support of the recommendation to group items in questionnaires. The current study tested the effect of randomly splitting a multi-item scale into two separate blocks of items that are positioned at the beginning and end of the same questionnaire. Our results indicate that this separation results in a significant drop in inter-item correlations for item pairs included in different blocks. When testing the factorial structure of the eight-item scale presented in two blocks of four items each, separated by a buffer of unrelated items, a clear and consistent pattern emerged: two internally consistent factors were extracted, and the pattern of internal consistency was completely in line with the way in which the items were grouped in the questionnaire. Moreover, the two factors clearly exhibited discriminant validity. This finding indicates that grouping items may lead to discriminant validity regardless of item content. Thus, evidence of discriminant validity that is obtained in grouped conditions should be interpreted with caution. Our results are consistent with findings showing higher inter-item correlations for items that are positioned in close proximity and/or on the same page (Tourangeau, Couper, & Conrad, 2004; Weijters et al., 2009). However, our study extends prior research in that it explicitly demonstrates the potential biasing effect of item grouping practices, especially with regard to factorial evidence for discriminant validity.

The observation that grouping items can lead to artificial discriminant validity raises important questions. Some implications and preliminary recommendations based on the

current study are as follows. First, testing the discriminant validity of constructs that are measured by scales positioned in different and clearly separated blocks in a questionnaire may stack the cards in favor of the anticipated multidimensional factor structure. As a consequence, the resulting evidence supporting discriminant validity should be interpreted with caution. Second, when constructing scales, researchers may need to experimentally test alternative ways of positioning items in a questionnaire. For example, when assessing the relation of a new scale with other constructs, researchers can compare the results based on two alternative formats: one in which the items of the new scale and its correlates are presented in random order in the same block, and one in which the new scale and each correlate are presented in separate blocks. Third, when reporting research results based on survey data, authors need to provide details on how the items were grouped and presented. If this information is missing, measures of (discriminant) validity and inter-construct correlations may not be comparable across studies. Fourth, it could be potentially misleading not to report all scales which were included in a questionnaire (including their positioning in the questionnaire), as their presence may have affected the results. And finally, in meta-analyses that assess relations between constructs, item positioning may be considered as a moderating variable.

The current study was not designed to investigate the psychological processes that lead to the patterns in the data (e.g., factorial structures), but recent research in the psychology of survey response provides some likely explanations. A first explanation relates to the information that respondents retrieve when answering survey questions. In line with belief sampling theory (Tourangeau, Rips, & Rasinski, 2000), Weijters et al. (2009) showed that the information retrieved in response to items that are positioned in close proximity to each other (and are not reversed) tends to show more overlap. With increasing physical inter-item distance, the overlap dissipates. A second explanation relates to stylistic responding. Evidence

has accumulated that response styles not only have a stable component (Billiet & Davidov, 2008; Weijters, Geuens, & Schillewaert, 2010b), but also a situational or local component that causes nearby items to be more highly correlated, regardless of shared or non-shared content (Weijters, Geuens, & Schillewaert, 2010a). In line with this, Gehlbach and Barge (2012) have suggested an anchoring and adjusting process that causes respondents to use their response to an item as an anchor for subsequent responses from which they insufficiently adjust. To illustrate, if respondents pick a four on a five-point rating scale for the first item, they may be biased toward providing a response close to four for the next item. Such anchoring is expected to occur for adjacent items that are similar to one another and for which the answer is at least somewhat uncertain (as is typically the case for attitudinal questions). If respondents engage in anchoring and adjusting, this inflates the correlations between adjacent items in a scale in the grouped format. But these inflated inter-item correlations are artificial and do not reflect a veridical high level of reliability (in the sense of shared true variance). In sum, our results do not provide insights into the psychological processes that cause these effects. Future research is needed to further clarify to what extent the results can be attributed to anchoring and adjustment (Gehlbach & Barge, 2012), the local component of response styles (Weijters, Geuens, et al., 2010a), and/or retrieval-related proximity effects (Weijters et al., 2009).

The current study has some other limitations, which offer avenues for future research on the topic. These limitations pertain to (1) the experimental design of the study, (2) the assessment of discriminant validity, and (3) the generalizability of our findings, in particular with regard to the scale format used. With regard to the experimental design, it would be useful to have a control condition in which the items were not grouped into separate blocks, as this would provide a more direct test of the effect of item blocking. The absence of a control condition in our design was mitigated by the use of a multi-level analysis of the observed correlations, in which we also included (as a covariate) the expected correlations as derived

from the scale development paper by Lastovicka et al. (1999). The results clearly indicate that, after accounting for the blocking manipulation in our data, the correlations in the current dataset were highly consistent with the correlations found by Lastovicka et al. (1999).

The grouping manipulation in the current study may have led to an increased probability of finding artificially created factors, as eight pages of four items each were positioned in between the two blocks of items from the same scale. It is not clear whether a similar result would have been obtained if the two blocks of same-scale items had been positioned in closer proximity, maybe even on two adjacent pages or screens. A further evaluation of the latter scenario would be particularly interesting as it may shed light on the question of whether or not survey designers should break up multi-item scales across more than one page or screen. With regard to the second limitation, the current study focused on discriminant validity in terms of finding changes in inter-item correlations and the emergence of artificial factors. Another important question is whether two factors that are obtained by splitting one scale into two blocks differentially predict a criterion variable of interest. This kind of assessment was not included in this study. Finally, we only employed one scale and a specific scale format, and survey researchers would probably like to see the phenomenon of artificial discriminant validity replicated across multiple constructs, across more than one study, and with alternative scale formats. With regard to the issue of scale formats, survey design experts have warned against the use of agreement rating scales as used in the current study (Artino, Gehlbach, & Durning, 2011; Converse & Presser, 1986; Fowler, 2009), although Likert-type scales are still widely employed in practice (Weijters, Cabooter, & Schillewaert, 2010). As mentioned by an anonymous reviewer, perhaps item formats that make respondents slow down to more fully and thoughtfully process items would not give rise to the biasing effect we observed. Further research may empirically test whether alternative formats are indeed less prone to the type of bias found in this study.

Table 1: Item descriptive statistics

Item statement	Condition 1 (<i>N</i> = 268)		Condition 2 (<i>N</i> = 255)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1. If you take good care of your possessions, you will definitely save money in the long run.	1.48	0.66	1.68	0.68
2. There are many things that are normally thrown away that are still quite useful.	1.86	0.87	2.06	0.81
3. Making better use of my resources makes me feel good.	1.62	0.72	1.65	0.75
4. If you can re-use an item you already have, there's no sense in buying something new.	1.82	0.80	1.81	0.84
5. I believe in being careful in how I spend my money.	1.81	0.84	1.67	0.77
6. I discipline myself to get the most from my money.	2.03	0.89	1.83	0.85
7. I am willing to wait on a purchase I want so that I can save money.	1.96	0.90	1.92	0.78
8. There are things I resist buying today so I can save for tomorrow.	2.07	0.92	2.03	0.77

Note: 1 = strongly agree, 5 = strongly disagree

Table 2: Observed correlations by condition and expected correlations based on Lastovicka et al. (1999)

Condition 1 ($N = 268$)								Condition 2 ($N = 255$)								Expected correlations								
i1	i2	i3	i4	i5	i6	i7	i8	i1	i2	i3	i4	i5	i6	i7	i8	i1	i2	i3	i4	i5	i6	i7	i8	
i1	0.51	0.67	0.53	0.44	0.38	0.45	0.34		0.24	0.39	0.22	0.31	0.27	0.47	0.45		0.47	0.49	0.49	0.53	0.43	0.52	0.50	
i2	0.51		0.51	0.48	0.23	0.24	0.34	0.29	0.24		0.25	0.16	0.15	0.12	0.20	0.19	0.47		0.44	0.44	0.47	0.39	0.46	0.45
i3	0.67	0.51		0.50	0.38	0.34	0.37	0.33	0.39	0.25		0.45	0.49	0.54	0.30	0.33	0.49	0.44		0.46	0.50	0.41	0.49	0.47
i4	0.53	0.48	0.50		0.33	0.31	0.45	0.39	0.22	0.16	0.45		0.46	0.42	0.31	0.23	0.49	0.44	0.46		0.50	0.41	0.48	0.47
i5	0.44	0.23	0.38	0.33		0.67	0.58	0.61	0.31	0.15	0.49	0.46		0.70	0.35	0.43	0.53	0.47	0.50	0.50		0.44	0.53	0.51
i6	0.38	0.24	0.34	0.31	0.67		0.55	0.55	0.27	0.12	0.54	0.42	0.70		0.34	0.36	0.43	0.39	0.41	0.41	0.44		0.43	0.42
i7	0.45	0.34	0.37	0.45	0.58	0.55		0.67	0.47	0.20	0.30	0.31	0.35	0.34		0.59	0.52	0.46	0.49	0.48	0.53	0.43		0.49
i8	0.34	0.29	0.33	0.39	0.61	0.55	0.67		0.45	0.19	0.33	0.23	0.43	0.36	0.59		0.50	0.45	0.47	0.47	0.51	0.42	0.49	

Note: Inter-item correlations by condition. Correlations for items that are in the same block are shown in boldface.

Table 3: Parameter estimates of the two-level model

Level	Parameter	Est.	90% CI		SE	Est./SE	p
			Lo	Hi			
Within level	β_1	-.165	-.198	-.131	.020	-8.06	<.001
	β_2	-.095	-.121	-.069	.016	-6.09	<.001
	$\text{var}(\varepsilon_{ic})$.003	.002	.005	.001	3.59	<.001
Between levels	γ_0	.145	-.131	.422	.168	.86	.387
	γ_1	.839	.272	1.405	.344	2.44	.015
	$\text{var}(\zeta_i)$.005	.002	.008	.002	2.84	.004

Est. = estimate; 90% CI = 90% confidence interval, with Lower boundary = Lo, Upper boundary = Hi. SE = standard error; p = p-value. The dependent variable in the model is r_{ic} (i.e., the observed correlation; see Table 2, conditions 1 and 2). For an explanation of the meaning of the various parameters, see the main text and the description surrounding equations (1) and (2).

Table 4: Fit statistics of alternative models

Condition	Model		χ^2	DF	CFI	TLI	SRMR	RMSEA
Condition 1	Two factors	Hypothesized model (see Figure 1)	53.82	19	.964	.946	.037	.083
		Best of alternative permutations	198.13	19	.810	.720	.090	.188
		Average of alternative permutations	221.84	19	.788	.687	.096	.200
	One factor	One-factor model	234.04	20	.776	.686	.095	.200
Condition 2	Two factors	Hypothesized model (see Figure 1)	43.67	19	.961	.943	.040	.071
		Best of alternative permutations	88.24	19	.880	.823	.069	.120
		Average of alternative permutations	129.93	19	.825	.743	.076	.151
	One factor	One-factor model	140.83	20	.810	.734	.076	.154

References

- Artino, A. R., Jr., Gehlbach, H., & Durning, S. J. (2011). AM Last Page: Avoiding five common pitfalls of survey design. *Academic Medicine: Journal Of The Association Of American Medical Colleges*, 86(10), 1327-1327.
- Billiet, J., & Davidov, E. (2008). Testing the Stability of an Acquiescence Style Factor Behind Two Interrelated Substantive Variables in a Panel Design. *Sociological Methods and Research*, 36(4), 542-562.
- Conte, J. M., Dean, M. A., Ringenbach, K. L., Moran, S. K., & Landy, F. J. (2005). The Relationship Between Work Attitudes and Job Analysis Ratings: Do Rating Scale Type and Task Discretion Matter? *Human Performance*, 18(1), 1-21.
- Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Beverly Hills: Sage Publications.
- Emons, W. H. M. (2009). Detection and Diagnosis of Person Misfit from Patterns of Summed Polytomous Item Scores. *Applied Psychological Measurement*, 33(8), 599-619.
- Fowler, F. J. (2009). *Survey research methods* (Fourth edition ed.). Thousand Oaks, CA: Sage Publications.
- Franke, G. H. (1997). "The Whole is More than the Sum of its Parts": The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires. *European Journal of Psychological Assessment*, 13(2), 67-74.
- Gehlbach, H., & Barge, S. (2012). Anchoring and Adjusting in Questionnaire Responses. *Basic and Applied Social Psychology*, 34, 417-433.
- Harrison, D. A., & McLaughlin, M. E. (1996). Structural Properties and Psychometric Qualities of Organizational Self-Reports: Field Tests of Connections Predicted by Cognitive Theory. *Journal of Management*, 22(2), 313-338.

- Hu, L. t., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.
- Lastovicka, J. L., Bettencourt, L. A., Hughner, R. S., & Kuntze, R. J. (1999). Lifestyle of the Tight and Frugal: Theory and Measurement. *Journal of Consumer Research*, 26, 85-98.
- Lehmann, D. R. (1988). An alternative procedure for assessing convergent and discriminant validity *Applied Psychological Measurement*, 12(4), 411-423.
- McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item Placement on a Personality Measure: Effects on Faking Behavior and Test Measurement Properties. *Journal of Personality Assessment*, 78(2), 348-369.
- Rick, S. I., Cryder, C. E., & Loewenstein, G. (2008). Tightwads and Spendthrifts. *Journal of Consumer Research*, 34(6), 767-782.
- Schriesheim, C. A. (1981a). The Effect of Grouping or Randomizing Items on Leniency Response Bias. *Educational and Psychological Measurement*, 41, 401-411.
- Schriesheim, C. A. (1981b). Leniency Effects on Convergent and Discriminant Validity for Grouped Questionnaire Items: A Further Investigations. *Educational and Psychological Measurement*, 41, 1093-1099.
- Schriesheim, C. A., & DeNisi, A. S. (1980). Item Presentation as an Influence on Questionnaire Validity: A Field Experiment. *Educational and Psychological Measurement*, 40, 175-182.
- Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989a). The Effect of Grouped Versus Randomized Questionnaire Format on Scale Reliability and Validity: A Three-Study Investigation. *Educational and Psychological Measurement*, 49(3), 487-508.

- Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989b). Grouped Versus Randomized Format: An Investigation of Scale Convergent and Discriminant Validity Using LISREL Confirmatory Factor Analysis. *Applied Psychological Measurement*, 13(1), 19-31.
- Shipp, A. J., Edwards, J. R., & Lambert, L. S. (2009). Conceptualization and measurement of temporal focus: The subjective experience of the past, present, and future. *Organizational behavior and human decision processes*, 110(1), 1-22.
- Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, Position, and Order. Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, 68(3), 368-393.
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*: Cambridge University Press.
- Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels. *International Journal of Research in Marketing*, 27, 236-247.
- Weijters, B., Geuens, M., & Schillewaert, N. (2009). The Proximity Effect: The Role of Interitem Distance on Reverse-Item Bias. *International Journal of Research in Marketing*, 26(1), 2-12.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The Individual Consistency of Acquiescence and Extreme Response Style in Self-Report Questionnaires. *Applied Psychological Measurement*, 34(2), 105-121.
- Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The Stability of Individual Response Styles. *Psychological Methods*, 15(1), 96-110.